

# Does universal 16S rRNA gene amplicon sequencing of environmental communities provide an accurate description of nitrifying guilds?

**Diwan, Vaibhav; Albrechtsen, Hans-Jørgen; Smets, Barth F.; Dechesne, Arnaud**

*Published in:*
Journal of Microbiological Methods

*Link to article, DOI:*
10.1016/j.mimet.2018.05.025

*Publication date:*
2018

*Document Version*
Peer reviewed version

# Does universal 16S rRNA gene amplicon sequencing of environmental communities provide an accurate description of nitrifying guilds?

Vaibhav Diwan, Hans-Jørgen Albrechtsen, Barth F. Smets, Arnaud Dechesne*

Department of Environmental Engineering, Technical University of Denmark, Miljoevej, 2800 Kgs, Lyngby, Denmark

Vaibhav Diwan, e-mail: vdiw@env.dtu.dk

Hans-Jørgen Albrechtsen, e-mail: hana@env.dtu.dk

Barth F. Smets, e-mail: bfsm@env.dtu.dk

* Corresponding author:

Arnaud Dechesne

Address: Department of Environmental Engineering, Technical University of Denmark, Building 115, Miljoevej, 2800 Kgs, Lyngby, Denmark

E-mail: arde@env.dtu.dk

Telephone: +45 45 25 22 91

Fax: +45 45 93 28 50

**Abstract**

Universal (i.e., targeting most bacteria/prokaryotes) 16S rRNA gene based amplicon sequencing is widely used for assessing microbial communities due to its low cost, time efficiency, and ability to provide a full overview of the community. However, it is currently unclear if it can yield reliable information on specific microbial guilds, which can be obtained by using primer sets targeting functional genes or specific16S rRNA gene sequences. Here, we compared the relative abundance, diversity, richness, and composition of selected guilds (nitrifiers), obtained from universal 16S rRNA gene based amplicon sequencing and from guild targeted approaches. The universal amplicon sequencing provided 1) accurate estimates of nitrifier composition, 2) clustering of the samples based on these compositions consistent with sample origin, 3) estimates of the relative abundance of the guilds correlated with those obtained from the targeted approaches and within ~1.2 orders of magnitude of them, but with measurable bias that should be considered when comparing estimates from both approaches. In contrast, the diversity and richness estimations using the universal 16S rRNA based amplicon sequencing were likely limited by the sequencing depth; therefore, we suggest preferring targeted approaches for assessing nitrifiers diversity and richness or using sequencing depth larger than those currently typically practiced.

Keywords: Diversity; Drinking water treatment; Wastewater treatment; Nitrifiers; AOB; *Nitrospira*.

## 1. Introduction

Community-wide high throughput amplicon sequencing of one or multiple hypervariable regions of the 16S rRNA gene has become a routine tool to describe the composition and diversity of microbial communities (Bartram et al., 2011; Caporaso et al., 2011). Beyond such an overall community assessment, it is often desirable to quantify and characterize specific constituent guilds in terms of abundance, composition, and diversity. For many guilds, the 16S rRNA gene can be informative because phylogenetic conservation of functional traits is common (Philippot et al., 2010; Martiny et al., 2013). In fact, if the microbial guild of interest consists of one or a few monophyletic clades (for example, ammonia-, nitrite-, or methane-oxidizers), it is possible to utilize 16S rRNA gene primers to specifically target the guild (Degrange and Bardin 1995; Kowalchuk *et al.*, 1997; Hermansson & Lindgren, 2001; Graham et al., 2007). Additionally, if conserved signature functional genes exist, targeting them can be a strategy. For example, *amoA*, which codes for a subunit of the ammonia monooxygenase can be used to target ammonia oxidizing prokaryotes; and *nxrB,* which codes for the beta subunit of the nitrite oxidoreductase can be used to target nitrite oxidizing bacteria (NOB) (Norton, 2011; Pester *et al.*, 2013). Targeting functional genes can be advantageous to access within-guild diversity, due to their high rate of evolution compared to the 16S rRNA gene (Dopheide et al. 2015).

In contrast to a non-specific approach such as universal (i.e., designed to target most bacteria) 16S rRNA gene amplicon sequencing, any guild-specific approach provides information only on the targeted guild but none on the rest of the community, which is an obvious limitation (Xue et al. 2013; Dopheide et al. 2015). The tradeoff is that universal approaches give a good overview of the whole

microbial community, but provide only limited information on non-dominant guilds because of their low contribution to the total sequence pool.

Therefore, it is currently unclear whether universal 16S rRNA gene amplicon sequencing is sufficient to obtain reliable information on specific microbial guilds and, more precisely, whether it correctly differentiates between samples with high and low guild diversity and whether it provides sufficient compositional information to identify samples with similar guild composition.

An additional concern regarding microbial community analysis using amplicon sequencing lies in its ability to provide reliable estimates of the relative abundance of the community constituents. Indeed, the quantitative nature of amplicon sequencing has been questioned (Zhou et al. 2011). Studies using mock communities have shown that the results from 16S rRNA gene based amplicon can be biased based on choice of primers and at different *in-vitro* (DNA extraction) and *in-silico* (sequencing and taxonomic assignment) experimental stages (Brooks et al. 2015; Parada et al. 2016; Thijs et al. 2017). Despite known biases, 16S rRNA gene based amplicon sequencing has been used successfully for quantitative assessment of the ubiquitous taxa in the bacterial community (Ibarbalz et al. 2014). However, to our knowledge, the literature provides no assessment of the reliability of universal amplicon sequencing for quantifying specific guilds. Therefore, one of the aims of this study was to fill this gap.

Here, using nitrifiers as model guilds, we compared bacterial 16S rRNA gene amplicon sequencing (further referred to as 'universal approach') to a guild-targeted approach in their abilities to infer relative abundance, diversity, richness, and composition. Nitrifying microbial guilds are a key to ammonium oxidation in many natural and engineered ecosystems. They were traditionally strictly divided into ammonia-oxidizing prokaryotes and nitrite oxidizing bacteria (NOB) before the recent

discovery of *Nitrospira* types that can fully oxidize ammonia (Comammox; Daims et al. 2015; van Kessel et al., 2015; Palomo et al. 2016). To target key nitrifying guilds, we used the functional genes *amoA* for ammonia oxidizing bacteria (AOB) and archaea (AOA); and *nxrB* for the main NOB genera *Nitrospira* and *Nitrobacter* (Arp and Stein 2003; Francis, Beman, and Kuypers 2007; Vanparys *et al.*, 2007; Pester *et al.*, 2013)

## 2. Material and Methods

### 2.1 Biomass sampling and DNA extraction

The biomass originated from the top (0-10 cm) layer of the after filter (AF) of 4 drinking water treatment plants (DWTP) in Denmark and from the nitrifying reactor (NR) of wastewater treatment plants (WWTP) from Denmark and Sweden and from an anammox reactor (Sjolunda (AR)) in a WWTP from Sweden (Figure S1). Three samples for each plant were used for qPCR. For the 16S rRNA gene amplicon sequencing, five samples were analyzed for three DWTP (Glostrup, Hillerod, and Odense (AF)) and single samples for all other plants.

DNA was extracted from 0.5 g of wet drained sand from DWTP and 0.5 g sludge from WWTP using FastDNA™ spin kit for soil (MP Biomedicals, Solon, OH, USA) according to manufacturer's instructions. Duplicate DNA extractions were done for each sample. DNA concentration was estimated using NanoDrop ND-1000 UV-VIS Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and the extracts were stored at -20 °C until further analysis.

### 2.2 qPCR, PCR and amplicon sequencing

Quantitative PCR (qPCR) analyses were conducted on a Chromo4 thermocycler using Opticon Monitor version 3 (BioRad). Each qPCR reaction contained 12.5 µL of 2X iQ SYBR green Supermix (Bio-Rad

Laboratories, Hercules, California, United States), 500 mM primer, DNA template (10 ng), and DNA/RNA-free water (Qiagen, Hilden, Germany) to 25µL. For each sample, DNA with average concentration ~26 and ~68 ng/µl for sand and sludge samples respectively was sent for 16S rRNA gene, *amoA*, and *nxrB* PCR, purification and amplicon sequencing (2 x 300 nucleotides) using the Illumina MiSeq platform at the DTU Multi Assay Core Center (Kgs. Lyngby, DK). For the universal 16S rRNA gene sequencing, primers (Bakt_341F and Bakt_805R, spanning the V3-V4 regions) and PCR conditions were from Herlemann *et al.*, 2011. For *amoA* AOB sequencing and qPCR, primers (amoA-1F and amoA-2R primers) and PCR conditions were from Rotthauwe et. al., 1997. For *nxrB Nitrospira* sequencing and qPCR, primers (nxrB169f and nxrB638r primers) and PCR conditions were from Pester *et al.*, 2013. Primers (CTO189fA-B, CTO189fC, and RT1r) and PCR conditions for AOB 16S rRNA gene qPCR were from Kowalchuk *et al.*, 1997. QPCR efficiencies and correlation coefficients obtained from the standard curves are in Table S1.

### 2.3 Bioinformatics analysis

Sequences generated as paired FASTQ files were processed using DADA2 (Version - 1.4; Callahan *et al.*, 2016)**.** DADA2 was preferred over other clustering-based methods as it recovers sequence variants that can vary by as little as a single nucleotide and thus avoids aggregation of sequences at arbitrary identity cutoff. DADA2 was used for quality filtering, trimming, de-replicating the reads, for inferring sequence variation through default error model parameters, for merging paired reads, for removing chimera, and for assigning taxonomy using Silva reference database v123 for 16S rRNA gene and custom taxonomy files for *amoA* and *nxrB*. A very similar pipeline was used for *amoA* and *nxrB*, except that for *amoA* only the forward reads (237 bp after trimming for quality) were analyzed due to the amplicon length. We applied frameshift correction (from package DECIPHER version 2.6; Wright

2016) and discarded sequences that were too distant from our reference sequences. the data obtained from DADA2 was analyzed using Phyloseq R Package (Version - 1.7.12; McMurdie and Holmes 2013). Raw sequence files were deposited into the sequence read archive at GenBank under the study accession number SRP127282.

### 2.4 Relative abundance estimation of AOB and *Nitrospira*

Universal 16S rRNA gene amplicon sequencing data was used to perform CaRcone analysis to obtain the average 16S rRNA gene copies per genome in the amplicon libraries (R script https://github.com/ardagulay/CaRcone---Community-average-rRNA-gene-copy-nr-estimator). The total numbers of bacteria per gram of biomass were derived by normalizing the gene copy number from qPCR with 1.75 16S rRNA gene copies per genome obtained from the CaRcone analysis. The abundance of AOB and *Nitrospira* cells per gram of biomass was obtained by assuming 2 *amoA* and 2 *nxrB* copies per genome (McTavish, Fuchs, and Hooper 1993; Lücker et al. 2010). Thus, comparing *nxrB* for *Nitrospira* and *amoA* for AOB with 16S rRNA gene was feasible as the abundance values obtained from two genes were normalized per genome.

The relative fractions of AOB and *Nitrospira* were calculated by dividing the normalized abundance estimates obtained from *amoA* and *nxrB* based qPCR by that obtained from 16S rRNA gene based total bacteria qPCR. Similarly, AOB and *Nitrospira* abundance estimates obtained from *amoA* and *nxrB* based sequencing were compared to their respective values obtained after dividing reads assigned to AOB and *Nitrospira* in the 16S rRNA gene based total bacteria sequencing by the total number of reads.

## 2.5 Statistical analysis

The alpha diversity metrics, Shannon diversity and observed richness were calculated using the 'estimate_richness' function in the Phyloseq R package version (Version - 1.7.12; McMurdie and Holmes, 2013). The estimates of Shannon diversity after rarefaction were calculated by performing rarefaction using the minimal number of sequences in each gene library, as we would lose a large amount of sequences from *amoA* and *nxrB* libraries if we rarified to the minimum number of sequences assigned to AOB and *Nitrospira* from the 16S rRNA gene library. For phylogenetic diversity calculation, first the sequences were aligned using the 'AlignSeqs' function in the DECIPHER R package (Version 2.0.2; Wright, 2015), then a neighbor joining (NJ) tree was constructed using phangorn R package (Version 2.3.1; Schliep, 2011), phylogenetic diversity (Faith's PD; the sum of the branches of the phylogenetic tree) was then calculated using the PhyloMeasures R package (Version 2.1; Faith, 1992; Tsirogiannis and Sandel, 2016).

To statistically compare the relative abundance and diversity estimates from universal and targeted approaches we used major axis (MA) regression (a Type II regression model) because for both type of estimates none of the variables (x and y) were experimentally controlled (Pierre Legendre and Legendre 1988). As the relative abundance estimates were homoscedastic on the log scale, we performed the regression analysis on the log scale with average abundance values from each plant. The MA regression was performed using 'lmodel2' package version 1.7-3 in R (P Legendre 2018).

To evaluate whether the universal and targeted approaches separated the samples similarly based on nitrifier composition, Principal Coordinate Analysis (PCoA) of samples based on their nitrifier composition obtained from universal and targeted approaches was performed separately. The PCoA

ordination matrix based on Bray Curtis similarities was generated using 'ordinate' function and the PCoA plot was generated using 'plot_ordination' function in the Phyloseq R package ( Version - 1.7.12; McMurdie and Holmes, 2013). Further, Procrustes analysis was performed to compare the multidimensional shapes of the two PCoAs by transforming them into a state of best superimposition (Peres-Neto and Jackson 2001). The significance of the similarity of the two PCoA matrices was then tested by a procrustean randomized test (PROTEST) that determines whether the sum of the residual nonconformities was less than that expected by chance (Jackson, 1995).

The clustering analysis was performed using the K-medoids clustering, a partitioning method that clusters a set of objects into K clusters, with K set based on a priori knowledge (Jin and Han 2010). K medoids clustering method was applied using PAM clustering in vegan (version 2.4-3) and cluster (version 2.0.6) R packages (Dixon 2003; Mächler *et al.*, 2012). The Clustering analysis figures were generated using 'clusplot' function in cluster R package (version 2.0.6; Mächler *et al.*, 2012) and further edited using Inkscape (version 0.921; Bah, 2007).

## 3. Results and discussion

We investigated biomass extracted from three nitrifying and one anammox wastewater treatment bioreactors and from four biological rapid sand filters producing potable water (Figure S1). At both types of plants, nitrifiers perform a fundamental role in ammonium removal and are known to range from 1% to 10% in abundance relative to the whole microbial community (Wagner et al., 2002; Gülay et al., 2016; Tatari et al., 2017).

In the universal approach, sequences assigned to nitrifier guilds were extracted to estimate the relative abundance, diversity, richness, and composition of nitrifying guilds (Table S2). These estimates were

then compared to those obtained from the guild-targeted approaches (qPCR for relative abundance and amplicon sequencing for diversity and composition; Table S2).

*Nitrobacter* (NOB) and ammonia-oxidizing archaea (AOA) were minor fractions of the nitrifiers in the analyzed communities. Indeed, both types of nitrifiers were undetected in the universal approach; the amplification of *Nitrobacter nxrB* prior to amplicon sequencing failed; and AOA were below or close to the detection limit of the *amoA* qPCR assay (10 gene copies per reaction) in WWTP and DWTP, respectively (data not shown). Therefore, here we focus only on AOB and *Nitrospira*; further collectively referred to as nitrifiers.

In the universal approach, 20,372 and 346,879 sequences were assigned to AOB and *Nitrospira*, respectively, across all sampling sites. From the targeted approach, 501,381 sequences were assigned to AOB (*amoA*) and 2,165,787 to *Nitrospira* (*nxrB*) (Table S3).

### 3.1 Relative abundance of nitrifiers

Targeted quantification of AOB was performed by qPCR using two primer sets targeting either betaproteobacterial-AOB-specific 16S rRNA gene or *amoA*. The relative abundances obtained from the universal approach were always lower than the 16S rRNA gene based qPCR (Figure 1A) but higher than the *amoA* qPCR (Figure S2-A). These results are consistent with previous observations from environments similar to those investigated in this study indicating that the 16S rRNA gene AOB primers from Kowalchuk et al., 1997, being unspecific, tend to overestimate AOB abundance, and that the *amoA* primers can underestimate AOB abundance depending upon the AOB composition (Figure S2-B; Dechesne et al. 2016; Tatari et al., 2017). Nitrifier composition for this study is presented and discussed in further sections. Compared to the universal approach, the *nxrB* qPCR yielded slightly higher relative abundance estimates for *Nitrospira* in all WWTP but lower for all DWTP (Figure 1B).

These observations for DWTP are consistent with previous findings of *Nirospira* relative abundance comparison of 16S rRNA gene based qPCR specific to *Nitrospira* and universal amplicon sequencing (Gülay et al. 2016). Overall, for both of the guilds, even if we observed inconsitstency between the targeted and the universal approaches, the universal approach based estimates were always within ~1.2 orders of magnitude of the targeted approaches.

Using MA regression analysis, we explored next whether the estimates from both approaches were linearly related. For AOB, the slope was significantly different from 0 (p value 0.003) and very close to 1 (slope estimate = 0.9), suggesting a direct proportionnality between the approaches. However, the intercept was negative (p value <0.05), consistent with the observation made erlier that the 16S rRNA gene based AOB quantification by qPCR tend to be significantly higher than the one based on universal amplicon sequencing (Figure 1-A). For *Nitrospira*, the intercept was not statistically different from 0 and the slope was significantly higher than 1 (slope estimate = 2.2) suggesting that the *Nitrospira* quantification based on universal amplicon sequencing increases more rapidly than *nxrB* qPCR with *Nitrospira* relative abundance (Figure 1-B).

Overall, the estimates of relative abundance of both guilds (AOB and *Nitrospira*) from universal approach correlated with those obatined from the targeted approaches and were within ~1.2 orders of magnitude of them, but with measurable bias. The universal approach certainly provides useful quantitative information for AOB and *Nitrospira* but the highlighted biases should be considered when comparing estimates from universal and targeted approaches for these guilds.

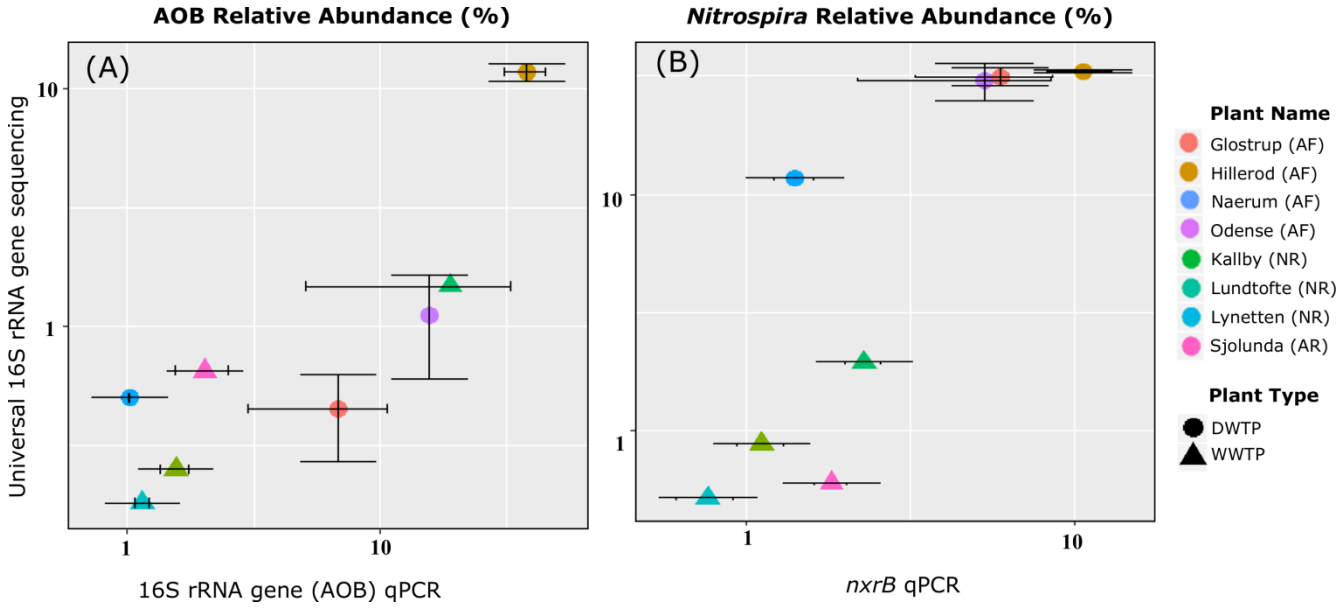


**Figure 1 (color in print):** Comparison of the relative abundance of AOB (A) and *Nitrospira* (B) based on universal (y-axis) and targeted (x-axis) approach. Universal 16S rRNA gene based amplicon sequencing was compared with qPCR (target: 16S rRNA gene for AOB in panel A and *nxrB* for *Nitrospira* in panel B) for four after filters (AF) from DWTP, three nitrifying reactors (NR) and one anammox reactor from WWTP. The error bars represent the standard deviation for each plant.

### 3.2 Alpha diversity of nitrifiers

For both guilds and irrespective of the metric used, the targeted approach always resulted in higher alpha diversity values compared to the universal approach (Figure 2). We ascribe this observation to the facts that the functional gene libraries contained approximately ~24 fold (AOB) and ~6 fold (*Nitrospira*) more sequences than retrieved from the universal sequencing (Table S3) and that diversity estimates can be highly reliant on the depth of sequencing (number of sequences per sample; Caporaso et al., 2011; Gihring et al., 2012; Smith and Peay, 2014). Additionally, the rate of evolution of functional genes (*amoA*, *nxrB*, *rpoB* etc.) is known to be higher than that of ribosomal genes (16S

rRNA gene), which makes them more phylogenetically resolutive ( Case et al., 2007; Pester et al., 2012; Pester *et al.*, 2013). Thus, using the targeted approach on functional genes results in the detection of more sequence variants than with the universal approach, which translates to higher but also, likely more accurate, diversity estimates.

For both guilds, the Shannon diversity and observed richness calculated from the universal approach correlated very poorly with that estimated from the targeted approach (Figure 2 A, B, D). This was also true when these indexes were calculated after rarefaction to the minimal number of sequences in each comparison (Figures S3 and S4). Phylogenetic diversity (Faith's PD; the sum of the branches of the phylogenetic tree) for AOB and observed richness (without rarefaction) for *Nitrospira* were the only diversity indices with a significantly positive correlation between both approaches (p value 0.009; Figure 2C and p value 0,01; Figure 2E). This frequent inconsistency between targeted and universal approach suggests that the guild-targeted approaches should be preferred to estimate the diversity of AOB and *Nitrospira*.

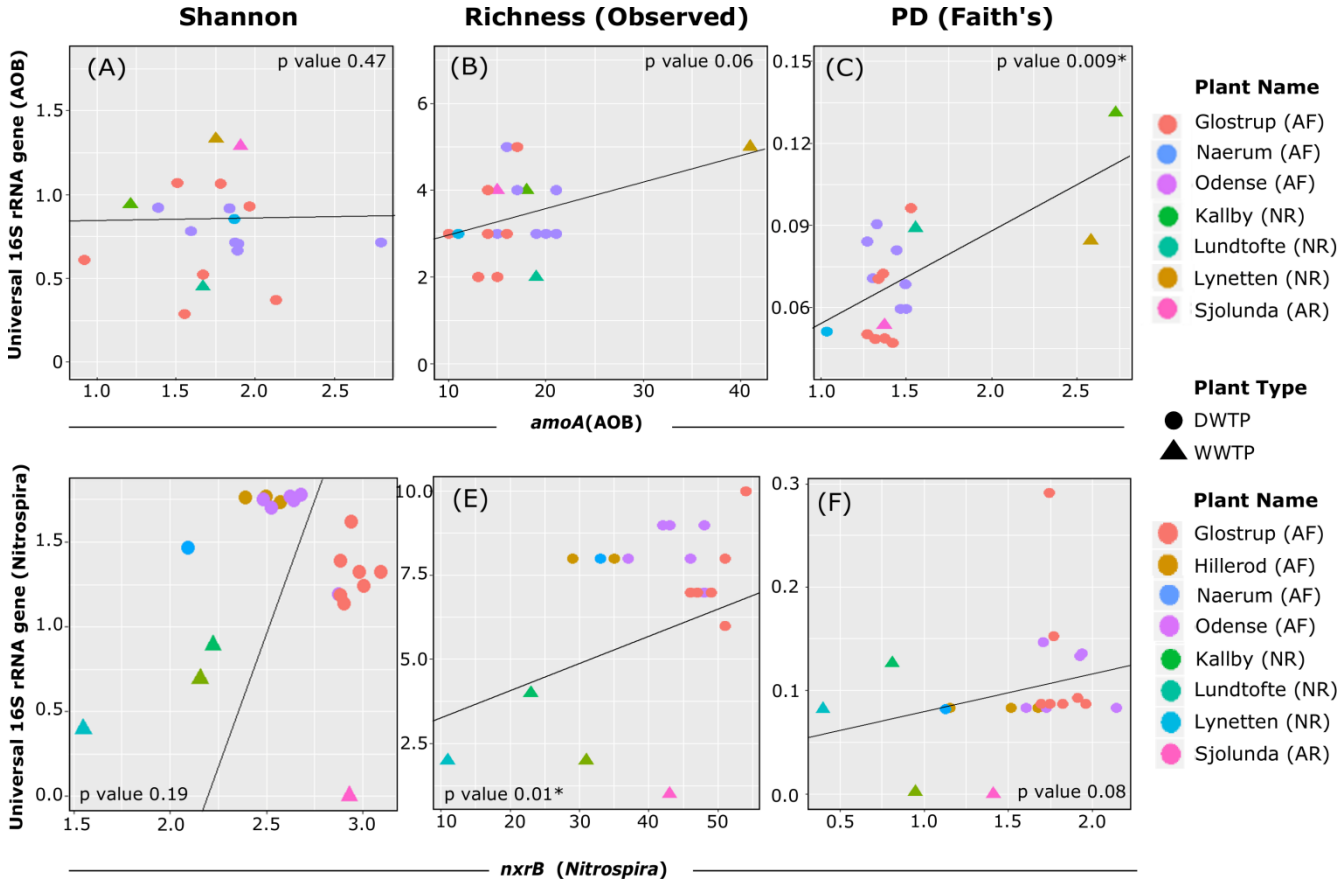


**Figure 2 (color in print):** Comparison of alpha diversity (Shannon; A and D), observed richness (B and E) and phylogenetic diversity (Faith's PD; C and F) of AOB and *Nitrospira* based on universal (16S rRNA gene) and targeted (*amoA* -top and *nxrB* -bottom) approaches. Scales on x and y-axis are different. The black line is the major axis regression and the respective p values are mentioned in each panel suffixed with a star sign if significant. For *Nitrospira*, a sample for which zero values were obtained with the universal approach (Sjolunda (AR)) was excluded from the linear regression. Likewise, for one DWTP, no valid AOB *amoA* sequences were obtained (Table S3).

## 3.3 Composition of nitrifiers

### 3.3.1 AOB composition

The universal approach only yielded sequences assigned to cluster 6 (represented by *Nitrosomonas aestuarii* and *Nitrosomonas marina*) and 7 (represented by *Nitrosomonas europaea* and *Nitrosomonas eutropha* strains) whereas the targeted approach also detected clusters 0, 2, 3, (*Nitrosospira*) and 8 (represented by *Nitrosomonas nitrosa*; Figure S5). Both approaches identified the same dominant cluster (Cluster 6) in five out of seven sites. Only for one plant (Sjolunda) did both approaches provide the exact same picture of the cluster-level composition. For all DWTP and one WWTP, the universal approach only showed the presence of cluster 6; whereas the targeted approach indicated the presence of additional clusters: cluster 7 in all these plants, plus cluster 0, 2, 3 and 8 in one WWTP (Kallby; (Figure S5). These observations relate with the low relative abundance of AOB observed earlier by universal 16S rRNA gene sequencing (Figure 1A) as clusters other than cluster 6 were largely unmapped for four out of seven sites by the universal approach due to their very low relative fraction to the total community.

Finally, at the subcluster taxonomic resolution, we found that all samples where cluster 6 was predominant consisted mainly of subcluster 6A members (represented by *Nitrosomonas oligotropha* and *Nitrosomonas* sp. ls79A3; Figure S6). As mentioned in previous studies, *amoA* AOB based primers (from Rotthauwe et al., 1997), having been designed primarily from cluster 7 members sequences, underestimate AOB abundance when subcluster 6A members are dominant (Dechesne et al., 2016). This likely explains why *amoA* based qPCR underestimated AOB in these plants (Figure S2).

### 3.3.2 *Nitrospira* composition

*Nitrospira* composition inferred from both approaches was largely similar for all sites (Figure S7). Both approaches identified strong compositional differences between DWTP and WWTP, the former being dominated by lineage 2 and the later by lineage 1 (Figure S7).

Overall, the universal approach successfully identified the dominant clusters/lineages within AOB and *Nitrospira* for most plants. Therefore, the universal approach can be a preferred choice for estimating the composition of these guilds, especially when the focus is on dominant guild members.

## 3.4 Samples ordination and clustering based on nitrifiers composition

In the PCoA analysis, both approaches separated DWTP and WWTP samples for both the guilds (Figure 3) but the PCoA plots obtained from the targeted approaches were visually different from that generated from their universal approach counterpart (Figure 3). Procrustean randomized test (PROTEST) was not significant for both guilds (AOB: p value = 0.8; *Nitrospira:* p value = 0.5), confirming the visual differences as significant.

Next, K-medoids clustering was performed to evaluate how the two approaches grouped (clustered) the samples based on their nitrifier composition. Here, we set K as the number of plants because we assumed that samples originating from the same plant should cluster together because communities from similar ecological sites tend to be similar in composition (Whittaker 1965).

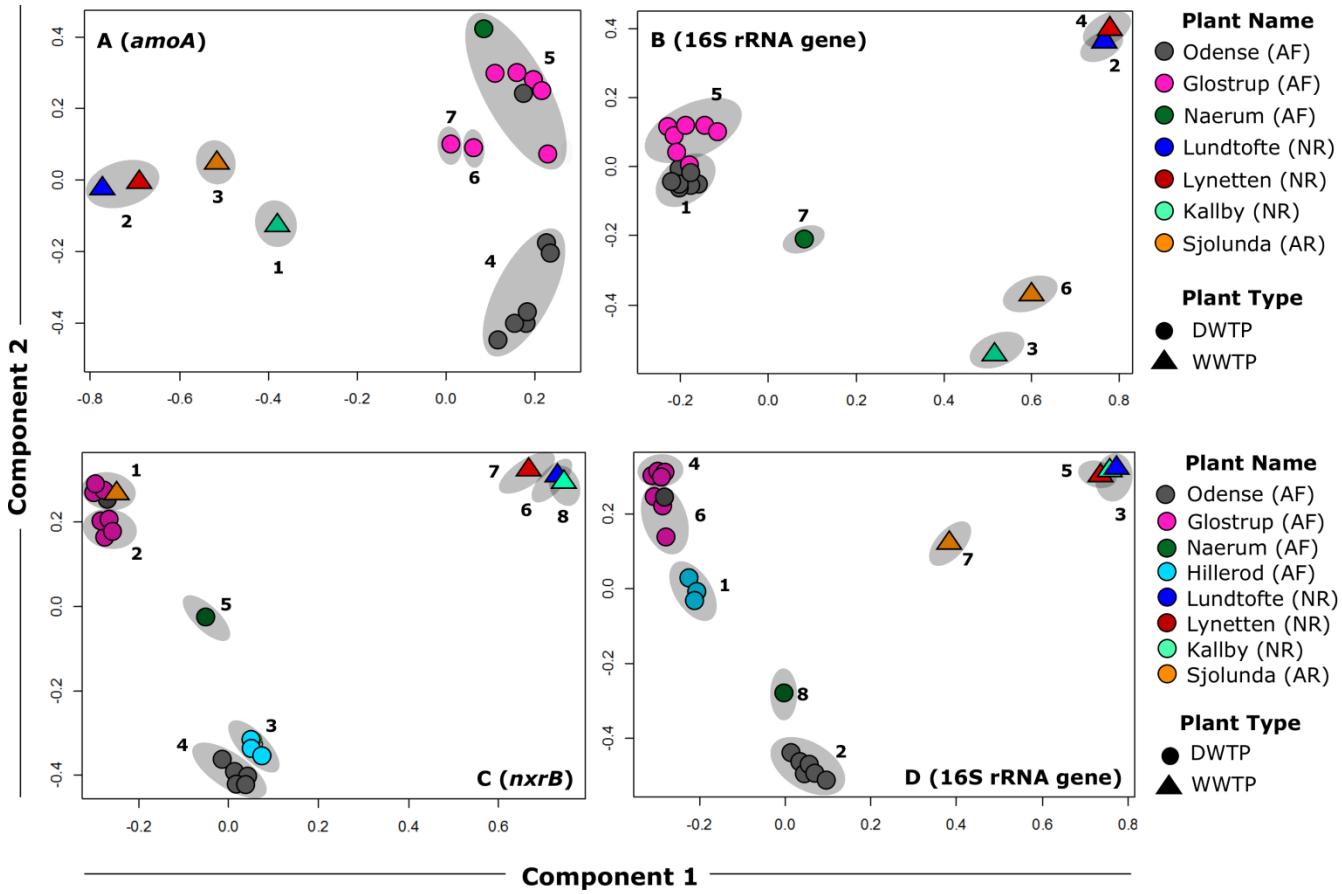


**Figure 3 (color in print):** Clustering analysis of samples after PCoA ordination of their compositions of AOB (panels A & B) and *Nitrospira* (panels C & D) obtained from the targeted and universal approach. Shaded ellipses followed by cluster number represent the clusters. Each solid circle and triangle is a sample, with color coding according to the plant. Seven replicate samples were used for Odense and Glostrup, three for Hillerod and one for all other plants. For AOB, the two components explain 50.4% of the variability for the PCoA based on *amoA* (A) and 57.3% for 16S rRNA gene (B). For *Nitrospira* the two components explain 67.5% of the variability for the PCoA based on *nxrB* (C) and 71.3% for 16S rRNA gene (D). The number of clusters (K) were adjusted to 7 and 8 for AOB and *Nitrospira*, respectively, according to the number of plants sampled.

### 3.4.1 Sample clustering based on AOB composition

The universal approach separated the samples into seven clusters corresponding to the seven plants, except for one replicate which was misplaced (Figure 3B). *amoA* incorrectly distributed the DWTP samples into four clusters, with one cluster comprising samples from all three plants (Figure 3A). Samples from two WWTP were placed in the same cluster (Cluster 2; figure 3A) whereas they were separated by the 16S rRNA gene (Cluster 2 and 4; figure 3B). Therefore, it is apparent that the targeted approach did not provide better clustering than the universal approach.

### 3.4.2 Sample clustering based on *Nitrospira* composition

The universal and targeted approaches for *Nitrospira* clustered samples similarly with some minor variations (Figure 3 C and D). For example, *nxrB* based analysis separated all plants except for one cluster which contained samples from two DWTP and one WWTP (Cluster 1; figure 3C). For the universal approach, all plants were separated except for one sample from a DWTP that clustered with samples from another DWTP (Cluster 6; figure 3D). Two WWTP that were separated (Cluster 6 and 8; figure 3C) by the targeted approach were now clustered together (Cluster 3; figure 3D). Here, the universal and the targeted approaches thus gave similar clustering outputs.

## 4. Conclusion

Taken together, our results provide the first systematic validation of the use of universal approaches for quantification and composition estimation of nitrifying guilds, as done by several authors (Rughöft et al., 2016; Ramanathan et al. 2017; Meerbergen et al., 2017; Saarenheimo et al. 2017). Overall, the 16S rRNA gene based universal amplicon sequencing of environmental communities can be efficiently used as substitute to more narrow targeted approaches for:

1) The assessment of nitrifier composition, as it captured the main ecological signal (the most dominant taxa) for all plants.

2) Sample clustering based on nitrifier composition, as it correctly separated samples from different plants.

The universal approach also provided relative abundance estimates within ~1.2 orders of magnitude of those from the targeted approaches, but with systematic biases that should be considered especially when comparing quantification estimates from both approaches.

Lastly, for the universal approach, diversity estimation was likely limited by the sequencing depth. Therefore, we suggest preferring targeted approaches for assessing nitrifiers diversity, unless much higher sequencing depths than commonly currently practiced are used.

**Conflicts of interest**

The authors declare no conflicts of interest.

**Acknowledgments**

We would like to thank Florian B. Wagner, Elena Torresi and Sara Ekström for assisting with sample collection. We also thank the plant operators at the Odense, Glostrup, Hillerod, Naerum DWTP and Lundtofte, Lynetten, Kally and Sjolunda WWTP for their cooperation.

**Funding:**

This research was financially supported by MERMAID (an initial training network funded by the People Programme - Marie Sklodowska-Curie Actions- of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 1607492).

## References


Arp, Daniel J., and Lisa Y. Stein. 2003. “Metabolism of Inorganic N Compounds by Ammonia-Oxidizing Bacteria.” *Critical Reviews in Biochemistry and Molecular Biology* 38 (6): 471–95. doi:10.1080/10409230390267446.

Bah, Tavmjong. 2007. *Inkscape: Guide to a Vector Drawing Program*. First. Upper Saddle River, NJ, USA: Prentice Hall Press.

Bansal, Arvind K, and Terrance E Meyer. 2002. “Evolutionary Analysis by Whole-Genome Comparisons Evolutionary Analysis by Whole-Genome Comparisons.” *Journal of Bacteriology* 184 (8): 2260–72. doi:10.1128/JB.184.8.2260-2272.2002.Updated.

Bartram, Andrea K., Michael D J Lynch, Jennifer C. Stearns, Gabriel Moreno-Hagelsieb, and Josh D. Neufeld. 2011. “Generation of Multimillion-Sequence 16S RRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads.” *Applied and Environmental Microbiology* 77 (11): 3846–52. doi:10.1128/AEM.02772-10.

Brooks, J Paul, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, et al. 2015. “The Truth about Metagenomics: Quantifying and Counteracting Bias in 16S RRNA Studies.” *BMC Microbiology* 15 (1): 66. doi:10.1186/s12866-015-0351-6.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” *Nat Meth* 13 (7): 581–83. http://dx.doi.org/10.1038/nmeth.3869.

Caporaso, J Gregory, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. 2011. “Global Patterns of 16S RRNA Diversity at a Depth of Millions of Sequences per Sample.” *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl (Supplement_1): 4516–22. doi:10.1073/pnas.1000080107.

Daims, H., E.V. Lebedeva, P. Pjevac, P. Han, C. Herbold, M. Albertsen, N. Jehmlich, et al. 2015. “Complete Nitrification by Nitrospira Bacteria.” *Nature* 528 (7583). doi:10.1038/nature16461.

Dechesne, Arnaud, Sanin Musovic, Alejandro Palomo, Vaibhav Diwan, and Barth F. Smets. 2016. “Underestimation of Ammonia-Oxidizing Bacteria Abundance by Amplification Bias in AmoA-Targeted QPCR.” *Microbial Biotechnology* 9 (4): 519–24. doi:10.1111/1751-7915.12366.

Degrange, V., and R. Bardin. 1995. “Detection and Counting of Nitrobacter Populations in Soil by PCR.” *Applied and Environmental Microbiology* 61 (6): 2093–98.

Dixon, Philip. 2003. “VEGAN, a Package of {R} Functions for Community Ecology.” *J Veg Sci 2003* 14 (6): 927–30.

Dopheide, Andrew, Gavin Lear, Zhili He, Jizhong Zhou, and Gillian D. Lewis. 2015. “Functional Gene Composition, Diversity and Redundancy in Microbial Stream Biofilm Communities.” *PLoS ONE* 10 (4): 1–21. doi:10.1371/journal.pone.0123179.

Faith, D P. 1992. “Conservation Evaluation and Phylogenetic Diversity.” *Biological Conservation* 61: 1–10. doi:10.1890/0012-9658(2006)87[1465:ATTFHF]2.0.CO;2.

Francis, Christopher A., J. Michael Beman, and M. M M Kuypers. 2007. “New Processes and Players

in the Nitrogen Cycle: The Microbial Ecology of Anaerobic and Archaeal Ammonia Oxidation." *ISME Journal* 1 (1): 19–27. doi:10.1038/ismej.2007.8.

Gihring, Thomas M., Stefan J. Green, and Christopher W. Schadt. 2012. "Massively Parallel RRNA Gene Sequencing Exacerbates the Potential for Biased Community Diversity Comparisons Due to Variable Library Sizes." *Environmental Microbiology* 14 (2): 285–90. doi:10.1111/j.1462-2920.2011.02550.x.

Graham, David W, Charles W Knapp, Erik S Van Vleck, Katie Bloor, Teresa B Lane, and Christopher E Graham. 2007. "Experimental Demonstration of Chaotic Instability in Biological Nitrification." *The ISME Journal* 1 (5): 385–93. doi:10.1038/ismej.2007.45.

Gülay, Arda, Sanin Musovic, Hans-Jørgen Albrechtsen, Waleed Abu Al-Soud, Søren J Sørensen, and Barth F Smets. 2016. "Ecological Patterns, Diversity and Core Taxa of Microbial Communities in Groundwater-Fed Rapid Gravity Filters." *The ISME Journal* 10 (9): 2209–22. doi:10.1038/ismej.2016.16.

Herlemann, Daniel P.R., Matthias Labrenz, Klaus Jürgens, Stefan Bertilsson, Joanna J Waniek, and Anders F Andersson. 2011. "Transitions in Bacterial Communities along the 2000 Km Salinity Gradient of the Baltic Sea." *ISME Journal* 5 (10): 1571–79. doi:10.1038/ismej.2011.41.

Hermansson, Anna, and Per-eric Lindgren. 2001. "Quantification of Ammonia-Oxidizing Bacteria in Arable Soil by Real-Time PCR" 67 (2): 972–76. doi:10.1128/AEM.67.2.972.

Ibarbalz, Federico M., María Victoria Pérez, Eva L M Figuerola, and Leonardo Erijman. 2014. "The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of

Bacterial Community Dynamics." *PLoS ONE* 9 (6). doi:10.1371/journal.pone.0099722.

Jackson, Donald A. 1995. "PROTEST: A PROcrustean Randomization TEST of Community Environment Concordance." *Écoscience* 2 (3): 297–303.

Jin, Xin, and Jiawei Han. 2010. "K-Medoids Clustering." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I Webb, 564–65. Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_426.

Kowalchuk, G A, J R Stephen, W De Boer, J I Prosser, T M Embley, and J W Woldendorp. 1997. "Analysis of Ammonia-Oxidizing Bacteria of the Beta Subdivision of the Class Proteobacteria in Coastal Sand Dunes by Denaturing Gradient Gel Electrophoresis and Sequencing of PCR-Amplified 16S Ribosomal DNA Fragments." *Applied and Environmental Microbiology* 63 (4): 1489–97. doi:10.1128/AEM.67.10.4880.

Legendre, P. 2018. "Lmodel2: Model II Regression." *R Package Version 1.7-3*. https://cran.r-project.org/web/packages/lmodel2/index.html.

Legendre, Pierre, and Louis Legendre. 1988. "Numerical Ecology, Volume 24." *(Developments in Environmental Modelling)* 24: 504. doi:10.1017/CBO9781107415324.004.

Liu, Zongzhi, Catherine Lozupone, Micah Hamady, Frederic D. Bushman, and Rob Knight. 2007. "Short Pyrosequencing Reads Suffice for Accurate Microbial Community Analysis." *Nucleic Acids Research* 35 (18). doi:10.1093/nar/gkm541.

Lücker, Sebastian, Michael Wagner, Frank Maixner, Eric Pelletier, Hanna Koch, Benoit Vacherie, Thomas Rattei, et al. 2010. "A Nitrospira Metagenome Illuminates the Physiology and Evolution

of Globally Important Nitrite-Oxidizing Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (30): 13479–84. doi:10.1073/pnas.1003860107.

Mächler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2012. *Cluster: Cluster Analysis Basics and Extensions. R Packages*. Vol. 1.

Martiny, Adam C, Kathleen Treseder, and Gordon Pusch. 2013. “Phylogenetic Conservatism of Functional Traits in Microorganisms.” *The ISME Journal* 7 (4). Nature Publishing Group: 830–38. doi:10.1038/ismej.2012.160.

McMurdie, Paul J, and Susan Holmes. 2013. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data.” *PLOS ONE* 8 (4). Public Library of Science: e61217. https://doi.org/10.1371/journal.pone.0061217.

McTavish, Hugh, James a. Fuchs, and Alan B. Hooper. 1993. “Sequence of the Gene Coding for Ammonia Monooxygenase\nin Nitrosomonas Europaea.” *Journal of Bacteriology* 175 (8): 2436–44. doi:citeulike-article-id:10421008.

Meerbergen, Ken, Maarten Van Geel, Michael Waud, Kris A. Willems, Raf Dewil, Jan Van Impe, Lise Appels, and Bart Lievens. 2017. “Assessing the Composition of Microbial Communities in Textile Wastewater Treatment Plants in Comparison with Municipal Wastewater Treatment Plants.” *MicrobiologyOpen* 6 (1): 1–13. doi:10.1002/mbo3.413.

Norton, Jeanette M. 2011. “Diversity and Environmental Distribution of Ammonia-Oxidizing Bacteria.” In *Nitrification*, edited by Martin G. Ward, Bess B. and Arp, Daniel J. and Klotz, 39–55. doi:doi:10.1128/9781555817145.ch3.

469 Palomo, Alejandro, S Jane Fowler, Arda Gülay, Simon Rasmussen, Thomas Sicheritz-Ponten, and
470 Barth F Smets. 2016. “Metagenomic Analysis of Rapid Gravity Sand Filter Microbial
471 Communities Suggests Novel Physiology of Nitrospira Spp.” *The ISME Journal* 10 (11): 2569–
472 81. doi:10.1038/ismej.2016.63.

473 Parada, Alma E., David M. Needham, and Jed A. Fuhrman. 2016. “Every Base Matters: Assessing
474 Small Subunit RRNA Primers for Marine Microbiomes with Mock Communities, Time Series and
475 Global Field Samples.” *Environmental Microbiology* 18 (5): 1403–14. doi:10.1111/1462-
476 2920.13023.

477 Peres-Neto, Pedro R., and Donald A. Jackson. 2001. “How Well Do Multivariate Data Sets Match?
478 The Advantages of a Procrustean Superimposition Approach over the Mantel Test.” *Oecologia*
479 129 (2): 169–78. doi:10.1007/s004420100720.

480 Pester, Michael, Frank Maixner, David Berry, Thomas Rattei, Hanna Koch, Sebastian Lücker, Boris
481 Nowka, et al. 2013. “NxrB Encoding the Beta Subunit of Nitrite Oxidoreductase as Functional and
482 Phylogenetic Marker for Nitrite-Oxidizing Nitrospira.” *Environmental Microbiology*, October.
483 doi:10.1111/1462-2920.12300.

484 Pester, Michael, Thomas Rattei, Stefan Flechl, Alexander Gröngröft, Andreas Richter, Jörg Overmann,
485 Barbara Reinhold-Hurek, Alexander Loy, and Michael Wagner. 2012. “AmoA-Based Consensus
486 Phylogeny of Ammonia-Oxidizing Archaea and Deep Sequencing of AmoA Genes from Soils of
487 Four Different Geographic Regions.” *Environmental Microbiology* 14 (2): 525–39.
488 doi:10.1111/j.1462-2920.2011.02666.x.

489 Philippot, Laurent, Siv G. E. Andersson, Tom J. Battin, James I. Prosser, Joshua P. Schimel, William

B. Whitman, and Sara Hallin. 2010. “The Ecological Coherence of High Bacterial Taxonomic Ranks.” *Nature Reviews Microbiology* 8 (7). Nature Publishing Group: 523–29. doi:10.1038/nrmicro2367.

Ramanathan, Bhargavi, Andrew M. Boddicker, Timberley M. Roane, and Annika C. Mosier. 2017. “Nitrifier Gene Abundance and Diversity in Sediments Impacted by Acid Mine Drainage.” *Frontiers in Microbiology* 8 (November): 1–16. doi:10.3389/fmicb.2017.02136.

Rotthauwe, J. H., Witzel, K. P., and Liesack, W. 1997. “The Ammonia Monooxygenase Structural Gene AmoA as a Functional Marker : Molecular Fine-Scale Analysis of Natural Ammonia-Oxidizing Populations.” *Applied and Environmental Microbiology* 63 (12): 4704–12.

Rotthauwe, J H, K P Witzel, and W Liesack. 1997. “The Ammonia Monooxygenase Structural Gene AmoA as a Functional Marker: Molecular Fine-Scale Analysis of Natural Ammonia-Oxidizing Populations.” *Applied and Environmental Microbiology* 63 (12). American Society for Microbiology: 4704–12.

Rughöft, Saskia, Martina Herrmann, Cassandre S. Lazar, Simone Cesarz, Shaun R. Levick, Susan E. Trumbore, and Kirsten Küsel. 2016. “Corrigendum: Community Composition and Abundance of Bacterial, Archaeal, and Nitrifying Populations in Savanna Soils on Contrasting Bedrock Material in Kruger National Park, South Africa [Front. Microbiol, (2016), 7, (1638)]. Doi: 10.3389/Fmicb.2016.01638.” *Frontiers in Microbiology* 7 (NOV): 1–16. doi:10.3389/fmicb.2016.01954.

Saarenheimo, Jatta, Sanni L. Aalto, Antti J. Rissanen, and Marja Tiirola. 2017. “Microbial Community Response on Wastewater Discharge in Boreal Lake Sediments.” *Frontiers in Microbiology* 8

(APR): 1–12. doi:10.3389/fmicb.2017.00750.

Schliep, Klaus Peter. 2011. “Phangorn: Phylogenetic Analysis in R.” *Bioinformatics* 27 (4): 592–93. doi:10.1093/bioinformatics/btq706.

Smith, Dylan P., and Kabir G. Peay. 2014. “Sequence Depth, Not PCR Replication, Improves Ecological Inference from next Generation DNA Sequencing.” *PLoS ONE* 9 (2). doi:10.1371/journal.pone.0090234.

Tatari, Karolina, Sanin Musovic, Arda Gülay, Arnaud Dechesne, Hans-Jørgen Albrechtsen, and Barth F. Smets. 2017. “Density and Distribution of Nitrifying Guilds in Rapid Sand Filters for Drinking Water Production: Dominance of Nitrospira Spp.” *Water Research* 127. Elsevier Ltd: 239–48. doi:10.1016/j.watres.2017.10.023.

Thijs, Sofie, Michiel Op De Beeck, Bram Beckers, Sascha Truyens, Vincent Stevens, Jonathan D. Van Hamme, Nele Weyens, and Jaco Vangronsveld. 2017. “Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S RRNA Gene Surveys.” *Frontiers in Microbiology* 8 (MAR): 1–15. doi:10.3389/fmicb.2017.00494.

Tsirogiannis, Constantinos, and Brody Sandel. 2016. “PhyloMeasures: A Package for Computing Phylogenetic Biodiversity Measures and Their Statistical Moments.” *Ecography* 39 (7): 709–14. doi:10.1111/ecog.01814.

van Kessel, Maartje A H J, Daan R Speth, Mads Albertsen, Per H Nielsen, Huub J M Op den Camp, Boran Kartal, Mike S M Jetten, and Sebastian Lücker. 2015. “Complete Nitrification by a Single Microorganism.” *Nature* 528 (7583). Nature Publishing Group: 555–59. doi:10.1038/nature16459.

Vanparys, Bram, Eva Spieck, Kim Heylen, Lieven Wittebolle, Joke Geets, Nico Boon, and Paul De Vos. 2007. “The Phylogeny of the Genus Nitrobacter Based on Comparative Rep-PCR, 16S RRNA and Nitrite Oxidoreductase Gene Sequence Analysis.” *Systematic and Applied Microbiology* 30 (4): 297–308. doi:10.1016/j.syapm.2006.11.006.

Wagner, Michael, Alexander Loy, Regina Nogueira, Ulrike Purkhold, Natuschka Lee, and Holger Daims. 2002. “Microbial Community Composition and Function in Wastewater Treatment Plants.” *Antonie van Leeuwenhoek* 81 (1–4): 665–80. doi:10.1023/a:1020586312170.

Whittaker, R H. 1965. “Dominance and Diversity in Land Plant Communities.” *Science* 147 (3655): 250 LP-260. http://science.sciencemag.org/content/147/3655/250.abstract.

Wright, Erik S. 2015. “DECIPHER: Harnessing Local Sequence Context to Improve Protein Multiple Sequence Alignment.” *BMC Bioinformatics* 16 (1). BMC Bioinformatics: 1–14. doi:10.1186/s12859-015-0749-z.

Wright, Erik S. 2016. “Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R.” *The R Journal* 8 (1): 352–59. doi:V12242009.

Xue, Kai, Liyou Wu, Ye Deng, Zhili He, Joy Van Nostrand, Philip G. Robertson, Thomas M. Schmidt, and Jizhong Zhou. 2013. “Functional Gene Differences in Soil Microbial Communities from Conventional, Low-Input, and Organic Farmlands.” *Applied and Environmental Microbiology* 79 (4): 1284–92. doi:10.1128/AEM.03393-12.

Zhou, Jizhong, Liyou Wu, Ye Deng, Xiaoyang Zhi, Yi-huei Jiang, Qichao Tu, Jianping Xie, Joy D Van Nostrand, Zhili He, and Yunfeng Yang. 2011. “Reproducibility and Quantitation of Amplicon

Sequencing-Based Detection.” *The ISME Journal* 5 (8). Nature Publishing Group: 1303–13. doi:10.1038/ismej.2011.11.